The Importance of Visual Attention in Improving the 3D-TV Viewing Experience: Overview and New Perspectives

Quan Huynh-Thu, Marcus Barkowsky, Patrick Le Callet

The Importance of Visual Attention in Improving the 3D-TV Viewing Experience: Overview and New Perspectives

Quan Huynh-Thu, *Member, IEEE,* Marcus Barkowsky, *Member, IEEE,* and Patrick Le Callet, *Member, IEEE*

***Abstract*—Three-dimensional video content has attracted much attention in both the cinema and television industries, because 3D is considered to be the next key feature that can significantly enhance the visual experience of viewers. However, one of the major challenges is the difficulty in providing high quality images that are comfortable to view and that also meet signal transmission requirements over a limited bandwidth for display on television screens. The different processing steps that are necessary in a 3D-TV delivery chain can all introduce artifacts that may create problems in terms of human visual perception. In this paper, we highlight the importance of considering 3D visual attention when addressing 3D human factors issues. We provide a review of the field of 3D visual attention, discuss the challenges in both the understanding and modeling of 3D visual attention, and provide guidance to researchers in this field. Finally, we identify perceptual issues generated during the various steps in a typical 3D-TV broadcasting delivery chain, review them and explain how consideration of 3D visual attention modeling can help improve the overall 3D viewing experience.**

***Index Terms*—3D-TV, perception, visual attention, human factors, video quality.**

I. INTRODUCTION

THREE-DIMENSIONAL (3D) content is receiving much attention as a result of a very strong push from the cinema industry. For 2011, more than fifty 3D movies have been announced for release [1]. With high-definition television (HDTV) being widely available now, the movie production companies and distributors, and the broadcasting and consumer electronics industries have been examining the next step that will enhance the television experience. As a result, 3D television (3D-TV) is being slowly deployed in the home environment.

Although there has been a rapid growth in 3D research, current literature has mainly focused on the technical challenges associated with the production, transmission and display of 3D-TV from the point of view of the broadcasters and content producers [2]. In this paper, we take a different view and discuss these issues from the point of view of the end-user and their perception of the 3D-TV content.

Depth perception is not produced solely by stereo-vision, as depth perception is possible with a single view provided it contains some monocular depth cues, e.g., motion, perspective, lighting, shading, occlusions. However, stereoscopic content provides important additional binocular cues that are used by human beings in the understanding of their surrounding world and in decision making.

The concept of 3D video has existed for a very long time but the latest advances in video technology and digital cinema (e.g., cameras, displays, optics, video processing) have made it possible to produce entertaining 3D content that can be viewed for an extended duration without necessarily causing extreme fatigue, visual strain and discomfort. However, the production of 3D stereoscopic content still represents a very difficult problem. Most existing solutions for content creation, production or post-production are still highly manual, expensive and time-consuming.

Besides artistic considerations related to content creation, broadcasters of 3D-TV programs have to overcome many engineering hurdles in the delivery chain: capture, storage, production, post-production, transmission, and display. Existing production workflows used in 2D cannot be easily used in 3D, even for the most traditional image manipulation (e.g., image scaling, rotation) and production tasks (e.g., insertion of logo or scrolling text, subtitles, transition effects).

Different formats are currently proposed for 3D-TV broadcasting, with future standards still to be agreed upon [3]. Furthermore, different 3D display technologies (e.g., LCD, PDP) and glass technologies (i.e., polarized, active-shutter) are currently available in the market. Standardized subjective testing protocols and objective (computational) tools to compare them in terms of visual perception and quality of experience would be very useful for researchers and for the industry. Some standardization bodies and standards-related groups have started to address these issues. ITU-R Working Party 6C is working towards the identification of requirements for the broadcasting and subjective testing of 3D-TV [4], while ITU-T Study Group 9 has recently added 3D video quality into its scope [5]. The Video Quality Experts Group (VQEG) 3D-TV ad-hoc group is examining the problem of reliable

Manuscript received July 15, 2010; revised February 28, 2011.

Q. Huynh-Thu is with Technicolor Research & Innovation, 1 Avenue de Belle Fontaine – CS 17616, 35576 Cesson-Sévigné, France (e-mail: quan.huynh-thu@technicolor.com).

M. Barkowsky and Patrick Le Callet are with the University of Nantes, France (email: {marcus.barkowsky,patrick.lecallet}@univ-nantes.fr).

subjective assessment of different quality aspects in 3D-TV.

A compelling visual experience by the end-user will be a key factor in the acceptance and use of 3D-TV. Most of the early 3D content is expected to come from movie releases and broadcast of live events. However, cinema and television viewing environments are very different, and perception of 3D content is therefore not similar in both cases. Adaptation of 3D content initially intended for a cinema screen is necessary for home viewing on a 3D-TV.

Studies have demonstrated that viewers tend to focus their attention on specific areas of interest in the image. Visual attention can therefore be considered a key aspect in determining the perception and overall visual experience. Models of visual attention have been proposed in the literature to automatically identify the main areas of interest in a picture. However, most of these works relate only to 2D video. Because the introduction of disparity information might affect the deployment of visual attention and because depth perception plays an important role on our attentive behavior when viewing 3D content, the understanding and modeling of 3D visual attention become relevant.

In this paper, we raise the importance of considering 3D visual attention when addressing 3D human factors issues. We provide a review of the state of the art in the field of 3D visual attention, discuss the challenges in both the understanding and modeling of 3D visual attention, and provide guidance to researchers in this field. Finally, we identify the perceptual issues caused by the different steps in a typical 3D-TV broadcasting delivery chain and discuss how the consideration of 3D visual attention modeling can be used to improve the overall 3D viewing experience.

The remainder of this paper is organized as follows. Section II describes the main human visual perception conflicts that can occur when viewing 3D-TV content and discusses their relationship with visual attention. Section III provides a review of advances in the understanding and the modeling of 3D visual attention. Section IV discusses the technical steps in the delivery of 3D-TV that may impact human visual perception, and provides solutions based on visual attention information to improve the overall viewing experience. Finally, we provide our conclusions in Section V.

II. RELATION BETWEEN PERCEPTION ISSUES AND VISUAL ATTENTION IN THE DELIVERY OF 3D-TV

Compared to existing (2D) TV services, 3D-TV stresses much more the visual perception of human viewers. Indeed, technological choices made to deploy 3D-TV are leading to some challenging consequences that the human visual system (HVS) has to deal with. A better identification and understanding of those effects are mandatory to improve the technology in order to make it more acceptable to end-users. In this section, we describe most of the important causes of stress that 3D-TV is raising from the point of view of the HVS. We discuss how visual attention consideration might be advantageous with a special focus on the accommodation-vergence conflict and binocular rivalry.

A. Depth Cues: Combination and Conflicts

One of the major goals of 3D-TV is to increase the sensation of presence or immersion through the enhancement of depth perception. This is mainly achieved by changing the binocular disparity that is related to binocular depth cues. Nevertheless, fortunately for stereo-blind observers, our visual system perceives depth using several cues, which are not only limited to binocular ones. Monocular cues such as occlusion of objects, perspective information, relative and absolute size of objects, motion parallax, accommodation, texture gradient, direction of the light source, and shadows complement binocular cues.

Depth cues have been analyzed and summarized concerning their accuracy and utility in [6]. All depth cues are fused together in an adaptive way depending on the viewing space conditions (personal, action, or vista space) but, to this date, only limited research has been done on the precedence, excitation and inhibition of the depth cues. In the Ponzo illusion it becomes evident that perspective provides a very strong depth cue, which supersedes for example the apparent size of known objects [7]. Studies on the relative importance of depth cues have also been conducted in the context of interactions with real-world objects in a virtual environment. Svarverud *et al.* have shown dependencies between stereoscopic disparity, motion parallax, and 2D-image based depth cues when locations of objects and distances are judged by humans in a virtual environment [8]. In some cases, adaptation is necessary as was shown for the case of accommodation by Bingham *et al.* [9].

When watching 3D content on currently available 3D-TV displays, depth cues might be in conflict. The strength of the conflict depends on the displayed content and in many cases the HVS can still correctly interpret the 3D scenario. One of the most apparent examples for this adaptation process is the switching of the views for the left and the right eye. While the binocular depth cues are opposite to all the other cues in this case, observers often report that they perceive stereoscopic 3D correctly. Nevertheless, after some time, they report visual discomfort, e.g., eye strain, headache or nausea. It might be assumed that the HVS is capable of performing a re-interpretation of the 3D depth cues but at a higher cognitive load. In general, enhancing the viewer's experience with binocular disparity might come as an additional workload on top of that required by monocular depth cues, especially if conflict is introduced. That is, a 3D-TV system might introduce cue conflicts that lead to visual discomfort.

Considering that visual attention might be helpful in limiting some of those conflicting effects. With current 3D displays, accommodation is quite limited as it is forced to be fixed on the display plane itself. Consequently, the natural defocus depth cue is rather poor and exhibits possible conflicts with binocular disparity. Knowing how visual attention is deployed on a given content might be useful to reintroduce defocus blur, through adaptive blurring, and thus limiting depth cue conflicts. On the other hand, this is surely steering visual attention itself, which could be seen as a limitation on observers' freedom to explore content. As a matter of fact, driving visual attention introducing retinal blur (according to

eccentricity) and defocus blur (according to depth) are candidate solutions to limit the cognitive load. In the next subsections, two other conflicts are detailed.

B. Vergence and Accommodation

Vergence and accommodation are mechanisms of the HVS that are closely linked and work together when the eyes view visual information in the real world. However, this is not the case with the display of 3D content on a television. On the one hand, as stated before, the eyes focus on the screen because objects appear sharpest on the display plane. On the other hand, the disparity of the objects between the left and the right eyes (or views) leads to a convergence of the eyes towards a point in front or behind the display plane, i.e., depicted objects can appear outside of the screen plane.

In the real world, an examined object is located at a certain position in space; therefore, the accommodation and vergence are in synchronization. When the human observer changes its attention to a different location in depth, both accommodation and vergence change at the same time. For the HVS, a change in accommodation induces an automatic change in vergence and vice-versa. The influence of blurred targets to the response of this coupling has been analyzed by Okada *et al.* [10].

When the human observer accommodates and converges on a certain point in depth, his/her vision is able to perceive objects within a certain range using stereopsis while objects farther away are not fused. The area in which fusion is possible is referred to as Panum's area and extends approximately 0.2-0.3 diopters. In order to avoid visual fatigue due to accommodation-vergence conflict, it has been suggested to restrict the displayed depth range to this region [11].

Currently, the extent of this area is modeled as a maximum and minimum allowed vergence in front and at the back of the screen. The consequences for displaying content from different sources and thus different disparities on typical 3D screens and the relationship to typical viewing distances has been shown recently in [12]. However, the size of the Panum's area depends on the extent of the targets, their spatial frequency, and the time of adaptation and a more sophisticated approach may be necessary. It also changes dramatically with the eccentricity as measured relative to the fovea. It has been reported that the fusion limit may be as small as 0.16 degrees in the fovea but reaches 0.5 degrees at 6 degrees eccentricity [13], [14]. We hypothesize that the consequence might be that a certain combination of objects can be fused in the peripheral vision but when it attracts attention later, it may happen that the observer is not able to fuse it anymore when in foveal vision. The simple assumption that the observer stays focused on the center of the screen may not be correct, in particular when large viewing angles and high-definition content is displayed. Concerning this specific conflict, it is intuitive to think that knowing the position of objects that have the potential to attract visual attention becomes very useful.

C. Binocular Fusion and Rivalry

In general, the HVS is able to align and fuse the views of the left and the right eye within certain limits at the gaze point. The limits and their consequences have been discussed in Section II-B. The exact mechanism of binocular fusion is still subject to controversial discussions and several different theories exist. A detailed review can be found in [15].

In the context of 3D-TV, it is important to note that the monocularly visible regions provide important insight on the structure of the scene [16]. These regions often indicate that an object that is close to the viewer occludes another object that is further away and thus can only be partly seen. At the edge of the two objects, only one eye can still perceive the partly occluded object. This becomes an issue in 3D-TV for example when those regions are extracted from view interpolation and no further information about the occluded object is available.

The binocular fusion is established even in the case of severe differences in terms of color, geometric distortions etc. For example, a red textured color plate presented to one eye and a green version presented to the other can still be fused [15]. In the 3D-TV scenario, this type of conflict occurs at several stages of the transmission chain. The television might display different colors or brightness levels to the two eyes. The video encoder might operate differently on the two views, e.g., when the bit rate controls of the left and the right view are not synchronized. As a consequence, blocky artifacts might be visible in one view but not in the other view. When the coding quality is very low, blocky artifacts might occur in both views. As the block structure is fixed on the same regular grid in both views regardless of the content, it appears as an overlay that is displaced in depth and not as a degradation of the objects. In the case of transmission errors, the concealment might be applied only in a single view or, in the worst case, the transmission error might lead to temporal de-synchronization where one eye is presented with temporally delayed content compared to the other eye.

In most of the cases discussed previously, the fusion succeeds but causes visual discomfort after a short period of time, presumably due to the additional cognitive load. Consequently, in the region of interest, it is beneficial to limit the conflicts due to binocular rivalry stemming from incorrectly reconstructed disoccluded regions or from coding and transmission errors. As in the 2D video case, a visual attention model can be useful to introduce hierarchy in the source content in order to apply adaptive processing such as unequal bit allocation or priority encoding. While in 2D, such techniques are supposed to benefit visual quality, in the 3D case their contribution might be even larger impacting also visual comfort, which constitutes with visual quality one important component of the quality of experience.

III. 3D VISUAL ATTENTION

A. Background

Research on visual attention modeling is nowadays at a cross-road between many different fields such as neuroscience, cognitive science, psychology, and image processing. Studies have indicated that viewers tend to focus their attention on specific areas of interest in the image and two mechanisms of

visual attention have been identified: bottom-up and top-down [17], [18]. Bottom-up attention relates to involuntary, automatic, and unconscious aspects of vision. It is mostly driven by signal characteristics. Top-down attention relates to voluntary and conscious aspects of vision. Eye-tracking experiments are conducted to study visual attention with two purposes: the recording of scan paths, usually represented or analyzed in terms of successions of fixations and saccades, and the identification of the locations of visual interest in the content (saliency). Models of visual attention are usually designed to produce (predict) saliency maps, which represent the location and level of visual interest of each area (or pixel) in the content. In this paper, we focus on the second aspect.

The idea of modeling visual attention for saliency prediction based on the use of visual features and integration theory appeared [19], and the idea of a biologically-plausible computational model followed [20]. Subsequently, models of visual attention have been proposed for various applications and have also been classified into three different categories: hierarchical models, statistical models and Bayesian models [21]. Most models proposed in the literature are based on a bottom-up architecture, often relying on the contrast detection of a number of low-level features such as luminance, color, orientation, motion, e.g., [22], [23], [24], [25], [26], and may use the concept of rarity [27] or surprise [28].

Research on visual attention modeling and its applications has increasingly gained popularity. However, compared to the amount of works on still images, relatively few studies have investigated visual attention modeling on moving sequences. Furthermore, only a very small number of works related to visual attention on stereoscopic 3D content can currently be found in the literature. However, this field has recently attracted interest because of the emergence of 3D video (in cinema and home) and recent availability of high-definition 3D-capable equipment to capture and display stereoscopic content. Depending on the philosophy or architecture used for modeling visual attention, the extension of existing 2D visual attention models to 3D content is not straightforward, especially in a biologically plausible way. Finally, collecting 3D gaze patterns and 3D saliency maps using existing eye-tracking equipment raises serious challenges.

B. Studies of Visual Attention in Stereoscopic Content

Although depth perception is possible with monoscopic images containing monocular depth cues, stereoscopic content brings important additional binocular cues enhancing our depth perception. A few studies have started to examine how visual attention may be influenced by such binocular depth cues.

Jansen *et al.* examined the influence of disparity on human behavior in visual inspection of 2D and 3D still images [29]. They recorded binocular data in a free-viewing task on 2D and 3D versions of natural, pink noise, and white noise images. Although eye position data were collected binocularly, analysis was performed using only the data from the left eye. The argument to use only data from the left eye was that the input to the left eye was identical over 2D and 3D versions of the stimuli, as the 2D version of the 3D image consisted of two copies of the left view. In order to investigate the role of disparity as a bottom-up process of visual attention, they selected visual stimuli showing only natural landscapes without any man-made objects. They investigated the saliency of several image features: mean luminance, luminance contrast, texture contrast, mean disparity (used as a measure for distance), and disparity contrast (used as a measure for depth discontinuity). The additional depth information led to an increased number of fixations, shorter and faster saccades, and increased spatial extent of exploration. The saliency of mean luminance, luminance contrast, and texture contrast was comparable in 2D and 3D stimuli. Mean disparity was found to have a time-dependent effect in 3D stimuli. Disparity contrast was found to be elevated only at fixated regions in 3D noise images but not in 3D natural images. They observed that participants fixated closer locations earlier than more distant locations in the image. Interestingly, they also found the same behavior in 2D images where depth perception was provided by monocular cues.

The study by Jansen *et al.* has shown that different depth cues have an influence on saccades. Based on these findings, Wismeijer *et al.* investigated if saccades are aligned with individual depth cues, or with a combination of depth cues [30]. In their experimental work, they presented subjects with surfaces inclined in depth, in which monocular perspective cues and binocular disparity cues specified different plane orientations, with different degrees of both small and large conflict between the two sets of cues. Additionally to recording eye movements, they asked participants to report their perception of plane orientation for each stimulus. They found that distributions of spontaneous saccade directions followed the same pattern of depth cue combination as perceived surface orientation: a weighted linear combination of cues for small conflicts, and cue dominance for large conflicts. They also examined the relationship between vergence and depth cues, and found that vergence is dominated only by binocular disparity.

Häkkinen *et al.* examined how stereoscopic presentation can affect eye movement patterns by presenting the 2D and 3D versions of the same video content [31]. Their results indicated that eye movements for 3D content are more widely distributed. They reported that observers did not only look at the main actors in the movie but eye movements were also distributed to include other targets. Their observations therefore corroborate those from Jansen *et al.* The study by Häkkinen *et al.* provided some interesting insights on the influence of the presentation of stereoscopic content on visual attention, showing that differences exist between 2D and 3D content. However in this study, eye movements and scan paths were only analyzed and discussed in the 2D sense by looking at the spatial spread of eye fixations in the image, without considering those aspects in terms of the gaze depth. Furthermore, in this study, participants were instructed to compare the two versions (2D vs. 3D) and to provide their opinion on which version they thought was better. It can be argued that this experiment was therefore driven by a quality judgment task and therefore involved top-down aspects of

visual attention.

As part of a study related to stereo-filmmaking, Ramasamy *et al.* [32] found that, in a scene showing a long deep hallway, gaze points were more spread in the non-stereoscopic version than the stereoscopic version as gaze points were more concentrated at the far end (in terms of depth) of the scene in the stereoscopic version. In other words, these results indicate that the spread of fixations could be more confined when viewing 3D stereoscopic content, and oppose the conclusions by Jansen *et al.* [29] and Häkkinen *et al.* [31].

A recent work has also examined the differences in visual attention between the viewing of 2D and 3D stereoscopic content [33]. In this study, twenty-one different video clips with a wide variety of spatio-temporal characteristics and range of disparity were shown to a panel of viewers in both their 2D and 3D stereoscopic version. Gaze locations were recorded using an eye-tracking equipment in a free-viewing task. Average saccade velocity was found to be higher when viewing 3D stereoscopic content, corroborating results from Jansen *et al.* [29] who used still images. However, other results in [33] did not corroborate those reported by Jansen *et al.*, as average fixation frequency and average fixation duration were found to be lower when viewing 3D stereoscopic content. Furthermore, the observations reported in [33] did not show evidence that fixations were more widespread when viewing 3D stereoscopic content, nor the opposite. The spread of fixations was found to be highly dependent on the content characteristics and narrative flow of the video, and not only on the depth effect provided by the binocular disparity. In a video with a single scene and static camera view, allowing viewers more time to explore different areas in the scene, fixations were more widespread in 3D than in 2D, suggesting that viewers do explore more the scene in that case. On the other hand, in a video with fast motion and many rapid scene changes, the spatial locations and densities of the fixations were very similar in both cases and often biased toward the center of the image. It was also found that specific content features carrying high cognitive information, such as text and faces, clearly attracted viewers' attention and therefore produced similar fixation patterns, regardless of the (2D or 3D stereoscopic) version. Nonetheless, differences in gaze patterns were found. Some background areas that did not clearly attract attention in 2D became in some cases areas of interest in 3D, especially in content providing sufficient time for viewers to explore the scene. Finally, the authors found that, even if fixation locations were similar in the viewing of 2D and 3D stereoscopic content, the temporal order of the fixated locations presented differences.

From the review of the different works above, we can conclude that the influence of the binocular depth cue on visual attention is highly dependent on the content itself and not only on the presence or strength of the disparity.

C. 3D Visual Attention Models

One early proposal of a computational model of depth-based visual attention for target detection came from Maki *et al.* [34], [35]. The simple architecture is based on a first stage of parallel detection of preattentive cues (image flow, stereo disparity, and motion detection), followed by a stage of cue integration using selection criteria based on nearness and motion. Two masks are first computed based on the pursuit and saccade modes found in human scan paths. Depth is then used to apply a priority criterion: in each frame either the target pursuit or the target mask is selected to be the final mask based on the depth. The hypothesis made by Maki *et al.* is that the target that is closer to the observer should be assigned highest priority. As discussed by the authors, this hypothesis may hold in a scenario where the observer has to avoid obstacles. However, we argue that this hypothesis may not necessarily hold in a scenario of viewing complex entertainment video content where the closest object may not be the only or main area of interest. In essence, the model proposed by Maki *et al.* serves only the purpose of detecting the closest moving object to the observer. Indeed, they demonstrated the application of their model using a scenario in which a moving or stationary stereo-camera selectively masks out different moving objects in real-scenes and holds gaze on them over some frames. They showed that their model kept focusing on the moving object that is the closest to the camera.

Ouerhani and Hügli also proposed a model of visual attention using scene depth information to extend a 2D saliency-based computational model [36]. Firstly, a number of low-level features are extracted from the image to build feature maps. Secondly, each feature map is transformed into a corresponding conspicuity map based on a multi-resolution center-surround mechanism. Finally, a linear combination of the conspicuity maps is used to produce an overall saliency map for the image. In order to integrate the effect of depth, additional depth-related features are extracted from the scene, resulting in additional conspicuity maps to be integrated in the final linear combination. Several depth features were initially considered: depth representing the distance from camera to observer, mean curvature providing information on the geometry of objects in the scene, and depth gradient providing information on depth changes in the scene. However, Ouerhani and Hügli only integrated depth as an additional feature in their proposed model. Using a few images, they showed the usefulness of integrating depth information in a computational model. However, there is no mention of formal subjective experiments using an eye-tracker to evaluate the performance of the model and the added value of depth in the model.

Although past works have studied the mechanisms of stereo-vision and have proposed perception models, one of the few studies on the modeling of stereo visual attention found in the literature were from Bruce and Tsotsos who discussed the issue of binocular rivalry occurring in stereo-vision and the deployment of attention in three-dimensional space [37]. They also discussed the difficult and biologically implausible translation of some types of 2D computational visual attention models to the case of stereo-vision. In particular, they singled out hierarchical models that extract basic and independent features to produce a saliency map used to predict shifts of attention. They criticized the independence of the extracted features as this is not completely biologically plausible and is

potentially more difficult to justify for the modeling of binocular visual attention. It is argued that stereoscopic visual attention models must take into account conflicts between the two eyes resulting from occlusions or large disparities. In other words, the behavior of each eye and it corresponding eye gaze cannot be considered independently.

There is a compelling argument that an appropriate model in a biological sense should accommodate shifts in the position of an attended event from one eye to another. Therefore, the representation of a 2D saliency map obtained independently for each stereoscopic view will discard the relationship and correspondence between the two eyes. Based on these considerations, Bruce and Tsotsos proposed a stereo attention framework from an existing 2D visual attention model using a visual pyramid processing architecture [38]. Their extension concerns the addition of interpretive neuronal units in the pyramid dedicated in achieving stereo-vision. The architecture of the model includes neurons tuned to a variety of disparities, while preserving the connectivity among interpretive units. The study by Bruce and Tsotsos used a few simple synthetic images with structural objects composed of mainly straight lines but did not mention any consideration on moving sequences with more complex scenarios. Unfortunately, results are not discussed in terms of comparison with ground-truth data collected through an eye-tracking experiment.

Despite the opinions of Bruce and Tsotsos on the applicability of hierarchical models to 3D video, Zhang *et al*. proposed a bottom-up visual attention model for stereoscopic content [39]. Their approach consists in extending a hierarchical model for the stereoscopic vision by using the depth map of the stereoscopic content as an additional cue. A spatial and a motion saliency map are constructed from features such as color, orientation and motion contrasts. A depth-based fusion with the spatial and motion saliency map is then used. The fusion is designed such that the relative importance of spatial or motion attention is set according to the strength of motion contrast.

We can raise two main criticisms concerning the work by Zhang *et al.* Firstly, one single spatial and one single motion saliency map are constructed from features extracted from either of the two views. Although the model is presented to handle stereoscopic video content, there is no mention of fusion of information between the two views or interview competition. The model seems more appropriate to a Multiview Video plus Depth (MVD) sequence, where each view has a corresponding depth map. Therefore the model can be considered as computing a prediction of visual attention independently for each view, rather than taking into account the stereoscopic perception of the video. Secondly, there is no mention of comparison of the model's results to ground-truth data obtained with human participants. It is therefore not possible to judge whether the addition of the depth cue had a significant impact on the prediction behavior of the original 2D bottom-up model.

D. Discussion on Issues and Challenges

Several research works, which have examined the influence of disparity and depth on visual attention, support the fact that depth provides a salient image feature. More importantly, these works provide some evidence that results of other past studies using 2D stimuli cannot be automatically generalized to 3D stimuli as the introduction of disparity information may change the deployment of visual attention.

A few computational models of 3D visual attention have been proposed. However, most of these works usually do not report results of formal subjective experiments to evaluate the performance of the proposed models. This raises an important question concerning the collection of ground-truth data of eye movements in the context of 3D-TV.

From the existing literature describing experimental work examining visual attention, it is apparent that there is a lack of standard protocols for conducting eye-tracking experiments in studies related to visual attention modeling and a lack of agreed procedures for post-hoc analysis of eye-tracking data. These issues already exist for the case of 2D video and still apply in the case of 3D.

Furthermore, additional hurdles related to the recording of binocular data and their interpretation in the 3D sense have to be considered in the study of 3D visual attention. Although some researchers have started to discuss modeling of 3D visual attention, the very first challenge that needs to be addressed is how to reliably collect and interpret ground-truth data. Most studies on 2D visual attention have been using monocular eye-trackers or binocular eye-trackers with the option of using the data from only one eye. In other words, past studies on visual attention have considered a similar behavior for each eye when recording eye movements of a participant viewing a 2D image or video. Although it is recognized that both eyes will not necessarily produce an identical gaze point on the screen when viewing a 2D content, this difference may not be relevant when considering a 2D saliency map. However, this difference of gaze location of each eye in a plane and in depth may prove to be more critical when trying to identify the 3D gaze point of both eyes.

For the recording of 3D gaze, binocular recordings are necessary. Yet, such eye-tracking equipment can only provide a two-dimensional spatial gaze location individually for each eye. These data then need to be extrapolated or processed to provide a notion of depth in relation with gaze direction or location [40], [41], [42]. The principles are similar to the ones involved in retinal disparity. By using two images of the same scene obtained from slightly different angles, it is possible to triangulate the distance to an object with a high degree of accuracy. If an object is far away, the disparity of that image falling on both retinas will be small. If the object is close or near, the disparity will be large. However, the triangulation of the two 2D gaze points from both eyes to produce a single 3D gaze point is not straightforward and is also dependent on the calibration of the system. For an experiment using 2D stimuli, calibration points are typically shown at different spatial locations on the screen and the observer is required to look at these points in order to calibrate the eye-tracker. In this case, it is easy to determine if the observer is looking accurately at the point since the 2D coordinates are known and the 2D gaze can

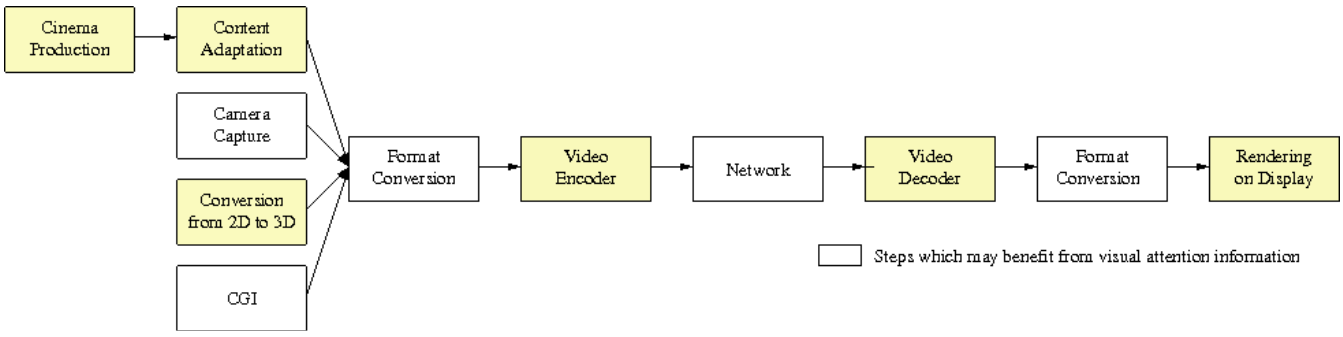

Fig. 1. 3D-TV delivery chain.

be accurately tracked on the plane of the screen. However, in the case of an experiment using 3D stimuli, this calibration procedure now requires a volumetric calibration, e.g., by showing points at different locations and different depth planes. Therefore, the problem is to find a procedure that ensures that the observer has accurately looked at the point at the given depth plane.

Past studies examining 2D visual attention on video tend to show that the HVS is more sensitive to motion contrasts compared to intensity or orientation contrasts. However it is not clear whether this is true in 3D. Further studies are required to examine visual attention in situations of opposing contrasts of motion, color and disparity. For example, where is visual attention drawn to in a scenario where motion contrasts are strong in the background but where there is a pop-out effect with a static object in the foreground?

Finally, existing studies on 3D visual attention have not discussed a 3D representation of saliency maps, as saliency maps are still currently considered purely in the 2D sense, i.e., saliency maps are still represented as flat maps.

IV. INTEGRATION OF VISUAL ATTENTION INFORMATION TO IMPROVE 3D-TV VIEWING EXPERIENCE

A. Overview

Each of the different processing steps in a typical 3D-TV delivery chain can introduce artifacts that may create issues for the human visual system. In this section, we briefly describe these processing steps and identify the related issues to the HVS. Furthermore, we focus on some of the 3D-TV processing steps by showing that integrating 3D visual attention information can be valuable in reducing the severity of visual artifacts, and in improving visual comfort and users' viewing experience. We provide various perspectives and discuss solutions, including their shortcoming, towards reaching these goals.

Fig. 1 shows a block diagram of a typical 3D-TV delivery chain, from content production to display at the end-user point. Several processing steps are required specifically for 3D-TV (e.g., 2D-to-3D conversion, format conversion). Furthermore, many of the steps that are normally present in a 2D-TV delivery chain (e.g., content production/capture, content adaptation) may introduce new issues in the 3D-TV scenario. Therefore, stereoscopic content can be affected by new sources/types of visual artifacts compared to 2D-TV.

B. 3D Video Capture

Multiple-camera setups require knowledge of the screen used for the final display in order to adequately adjust the disparity. If the display device changes (for example, digital cinema content that will be displayed on a television screen in the home environment), the disparity needs to be adjusted for the new screen size [12]. This has important implications with the rendering and depth perception as discussed in more detail in Section IV-D.

In the acquisition of stereo content, often two separate cameras are used, thus resulting directly in a stereo pair. Depending on the precision of the alignment, several artifacts can be introduced such as vertical misalignment, color misalignment, different focus points or zoom levels, or temporal offsets. Depending on the configuration of the stereo-cameras (i.e., parallel or toed-in), different corrections of geometry distortions are necessary (i.e., correction of keystone effect for the toed-in configuration and image shift for the parallel configuration). If the stereo images are not properly corrected, the visual artifacts will often result in visual discomfort, or in the worst case can result in the impossibility to correctly fuse the images.

Stereo-filmmaking cannot use the same shooting techniques that are used in 2D because some transitions or effects used in 2D do not work well in 3D. Conversely, 3D content production may lead to new ways of shooting or presenting content. New constraints also need to be considered to avoid situations where viewers are unable to focus or fuse the stereo-images.

A production tool that could provide some insights on the audience's attentional behavior would be valuable for stereographers. Indeed, such a tool would be useful in evaluating how viewers react to a new shooting technique by identifying the elements of interest or distraction in the scene. A 3D visual attention model would advantageously replace the need to conduct experiments a posteriori and could be used on-site.

C. 2D-to-3D Conversion

The amount of 3D content is currently still very limited

because stereoscopic filming is technically very challenging, requires dedicated equipment and experienced stereographers. Since a vast amount of 2D content is available, the industry is highly interested in the possibility to convert existing or new 2D video content to 3D.

However, this 2D-to-3D conversion is also technically very difficult, currently highly manual and time-consuming. Television manufacturers have started to introduce an automatic 2D-to-3D conversion functionality in their 3D-TV sets but results are currently far from being exempt of visible artifacts. Most of the algorithms for automatic 2D-to-3D conversion use a succession of the following steps: scene segmentation into objects, depth map generation, parallax computation, and 3D image construction (for example using pixel shifting and inpainting).

The generation of the depth map relies heavily on the existence of usable monocular depth cues in the 2D content. However, 2D content may not always contain enough of such depth cues for conversion. To overcome this limitation, the use of a saliency map computed from visual attention analysis has been proposed to replace the depth map in the automatic conversion process [43]. In this case, the saliency maps is used as input to the parallax computation on the hypothesis that salient regions are nearer to the observer and non-salient regions farther from the observer.

The limitation of this proposal is that it is based on the assumption that the areas of interest are always placed in the foreground of a scene. This is not always correct as discussed in Section III-B. Nonetheless, this work illustrates that visual attention analysis can help 2D-to-3D conversion. We suggest that, instead of replacing the depth map generation, saliency maps could be combined with depth maps to improve the results from automatic 2D-to-3D conversion.

D. Content Repurposing and Depth Adaptation

Content repurposing (also termed reframing) of 2D video content is used to address the problem of aspect ratio difference between the content and the screen (e.g., content with cinema aspect ratio shown on a display with a different aspect ratio). Without content repurposing, either black borders have to be inserted around the image (top/bottom or left/right) to fit the aspect ratio of the new target screen or geometric distortion will have to be applied on the content to fit the screen. Either approach will worsen the visual experience.

To adapt 2D video content to a display with a different aspect ratio, content repurposing usually involves a combination of cropping and zooming (re-scaling), especially for viewing on small devices. In this case, cropping is done around a region of interest selected to preserve the most important information in the content, i.e., the most visually important areas. Currently, determining the coordinates of the cropping window for content repurposing is a highly manual process, which can prove to be very time-consuming and expensive, or is performed simply to retain the center of the picture regardless of the content. A visual attention model could be used to make this process faster, more automated, and adaptive to the content as it would analyze the content to predict the main area of interest to retain in the reframing process. Such tool would be particularly very useful for the broadcasting of live events. Attention-based video reframing has already been proposed in the literature for 2D video [44], [45], [46] but a similar technology would be useful in the framework of 3D-TV broadcasting.

However, repurposing of 3D content needs to address two additional important issues. The first one is the border effect, which needs to be avoided as cutting objects appearing in front of the screen inhibits the perception of depth. The second one is depth adaptation. Creation of 3D stereoscopic content cannot be disconnected from the display and viewing conditions because both depth perception and visual comfort are highly dependent on the screen size and viewing distance, for a given content disparity. Therefore, a given content is currently produced for a given set of screen size and viewing distance. However, these factors are extremely different in cinema, home television, and on a portable device. Therefore, depth adaptation of stereoscopic content initially produced for a given viewing scenario is necessary for usage in a different one.

Content adaptation from cinema environment with its large field of view to the home environment with a narrower viewing angle and shorter viewing distance is currently technically very challenging. It is also a very time-consuming and manual process. More automated content-based post-production or post-processing tools to help adapt 3D content to television are required. Again, a 3D visual attention model would provide the area of interest and convergence plane to drive the content repurposing of stereoscopic content.

In addition to the necessary depth adaptation for 3D content repurposing, the adaptation of the scene depth can be used in order to improve visual comfort. The adaptive rendering of a 3D stereoscopic video based on identification of a main region of interest has been proposed using a 2D visual attention model [47]. The adaptation of the convergence plane of the main area of interest is used to reduce visual fatigue induced by a continuous need to change the plane of convergence when the main area of interest is moving across different depth levels. A way to reduce such strain is to modify the convergence plane of the main area of interest to place it on the screen plane, i.e., by adapting the content disparity. In order to achieve this, two steps are used. Firstly, a visual attention model is used to compute the saliency maps that indicate the importance of visual interest of all pixels in each view. Secondly, disparity information (depth or disparity map) is used to refine these saliency maps in order to select one dominant area of interest. The disparity information is also used to compute the necessary shift in disparity that has to be applied between the stereo-views to place the main area of interest on the zero-parallax plane. A succession of shifting, cropping and scaling is necessary to achieve this.

The drawback of the proposed approach is that cropping may introduce border effects if an object appearing in front of the screen and initially placed very near a border is cut in the reframing process. Scaling may also slightly deform the objects. The effectiveness of this approach also depends on the

availability and quality of the disparity map. The computation of high-quality disparity maps is still a difficult problem to be solved. Therefore, the approach based on the combination of computed disparity maps and a 2D visual attention model may suffer from the errors in these disparity maps. A proper 3D visual attention model may overcome this drawback.

E. Encoding of 3D Video

Numerous approaches exist to encode and transmit 3D video signals. The easiest is a simulcast transmission of the different views or depth maps using standard 2D video codecs such as H.264/AVC [48]. An extension to H.264/AVC, called Multiview Video Coding (MVC), was developed to allow the compression, transmission and storage of 3D video. MVC was adopted as a standard format on the Blu-ray Disc.

The independent or combined transmission of 3D video signals leads to new artifacts which most often lead to binocular rivalry, as discussed in Section II-C. Moreover, each compression algorithm requires a specific input representation, thus conversions between formats frequently occur, leading to information loss.

Video encoding that uses different compression parameters in the regions of interest (ROI) in the content has been proposed for 2D video, e.g., [49]. Since the problem of video compression is essentially the same in 2D and 3D, i.e., fit a sparser representation of a signal into a limited bandwidth, we can foresee that similar ROI-based compression can be applied in the context of 3D video encoding.

F. Decoding and Rendering of 3D Video

At the display side, another format conversion may occur depending on the signal representation used for transmission or if a different viewpoint needs to be rendered.

Depth Image Based Rendering (DIBR) approaches that rely on depth/disparity maps are frequently used. These render the stereo pair before the display, producing a dedicated image for the left and the right eye. Because at least one viewpoint differs slightly from the transmitted view, inpainting algorithms are needed to fill the previously occluded image regions. The inaccuracy of the inpainting often produces artifacts around the edges of objects.

3D rendering may require an estimation of the depth or disparity. Estimation of depth from at least two views is likely to produce artifacts, mostly because of the ambiguity of image features.

Existing 2D decoders usually employ some kind of error concealment techniques or freeze the last correctly decoded video frame when transmission errors occur. In the case of stereoscopic video transmitted as two streams, there is also the crucial issue of keeping the streams completely synchronized, especially if the decoder employs the frame freezing strategy when transmission errors occur. Indeed, a strong effect of binocular rivalry occurs in the case of simulcast transmission if the error concealment method leads to a temporal de-synchronization: the two stereoscopic images would belong to different object or camera positions and thus fusion might be difficult or impossible to achieve.

Besides temporal de-synchronization, spatial error concealment strategies may have a major impact on the perceived quality. In the 2D transmission case, error concealment methods are applied to predict the content of the missing image regions using spatial and temporal information available from the bitstream. Recently, it has been shown that such an error concealment strategy does not necessarily improve the quality of experience (QoE) for 3D videos in the same way. For transmission outages that affect the content over more than a few frames, switching back to 2D presentation seems to be preferred to either concealing the erroneous frames or to pausing the playback while staying in 3D presentation mode [50]. Another study subjectively evaluated the quality drop due to frames lost frequently, for example every other frame was lost in one view. In this case, it seems better to stay in the 3D presentation mode and to pause for one to three frames rather than switching to 2D for a single correctly received frame every two to four frames [51].

It may be anticipated that the switching between 2D and 3D presentation mode has an important impact on the annoyance of the viewer. This effect may be limited by determining the main region of interest and aligning the corresponding object to the display plane. Another alternative would be to apply an adapted error concealment method that conceals the erroneous regions in 2D or 3D in function of their visual importance.

G. Increase of Visual Comfort with Blurring Effects

Comfortable viewing conditions, i.e., zone of comfortable viewing, of stereoscopic content is linked to several factors among which range of depth of focus and range of fusion [52], [53].

For instance, Wöpking's studies [54] suggest that visual discomfort increases with high spatial frequencies and disparities. This is partially explained by the fact that the limits of fusion increase as a result of the decreased spatial frequency. More generally, it appears that blurring effects can positively impact visual comfort because they reduce the accommodation-vergence conflict limiting both accommodation and effort to fuse [55], [56].

As the distance from the fixation point increases, objects are perceived more and more blurred. Blur in this sense may be used to mimic the perception of retinal defocus, which will lead to the sought positive effect in visual comfort. Simulating depth-of-field (DOF) is a way to take advantage of this property, by artificially blurring images to a degree that corresponds to the relative depth or distance from fixated objects. As reported by Lambooij *et al.* [57], "three essential steps are required for proper implementation of a simulated DOF: localization of the eye positions, determination of the fixation point and implementation of blur filters to non-fixated layers". This procedure has been applied in virtual reality environment but is still subject to some drawbacks in more general contexts (depth cues integration between retinal disparity and high amount of blur [58]).

Blurring effects can also be used in 3D content to direct the viewer's attention towards a specific area of the image that should fit ideally in a zone of comfortable viewing. Although gaming is not a topic of interest in this paper, we can however

mention that visual attention modeling has attracted a growing interest from the computer graphics community, especially for virtual environments. The use of visual attention models has been proposed to produce a more realistic behavior of a virtual character, to improve interactivity in 3D virtual environments and to improve the visual comfort of the rendering of 3D virtual environments [59], [60], [61].

H. 3D Subtitling

Captions or subtitles may need to be inserted into 3D content and this needs to be done coherently, taking into account the possible problems of occlusions. In addition, a large difference between the convergence plane of the subtitles and that of the content of interest can lead to difficulties in viewing both simultaneously. For cinema viewing, production studios may choose the option of translating a movie in different languages to avoid the problem of insertion of subtitles in the 3D movie. However, in a home environment (e.g., Blu-ray disc or broadcast), subtitling remains an option for viewers as they may choose to watch a program in the original language with subtitles. Currently, the insertion of subtitles in a 3D content is a time-consuming manual process as subtitles are inserted by an operator frame by frame. Not only should the subtitles be spatially placed adequately in the video frame but they also must be placed at a suitable depth to minimize eye strain and avoid visual conflict with objects in the image.

For 3D subtitling, one approach is to always place subtitles in the screen plane (zero parallax) but this may generate visual discomfort due to occlusion by the content and other areas of interest present at a different depth in front. Another approach is to always insert the subtitles in front of the object closest to the viewer but this may create extreme disparities in front of the screen which are difficult to fuse and therefore create visual discomfort. Extreme visual fatigue will also likely be caused by the need for the viewer to always switch between accommodation planes to read the subtitles and look at objects of interest that could be at a different depth plane. Finally, a third approach is to shift the disparity between the views in order to move the region with largest negative disparity on the display plane (zero parallax). As a consequence all the 3D effect is confined inside/behind the screen and the subtitles can be inserted on the display plane (zero parallax). This would reduce visual fatigue as it is easier for human eyes to accommodate and converge behind the screen. However, the 3D dynamics of the scene is completely modified by such process.

The alternative strategy that could solve all the mentioned drawbacks would be to use depth information in the scene (e.g., extracted from the stereo pairs) and use a depth-dependent subtitle placement based on the convergence plane of the main area of interest in the image, which could be predicted using a 3D visual attention model.

V. CONCLUSION

The production and transmission of 3D-TV brings new types of visual artifacts and specific issues in perception that did not exist for 2D program content. Viewing conditions are very different in a home environment compared to cinema theaters and, for a given content, the different conditions can create a different perception of depth for the audience. Human perception is an important aspect that needs to be taken into account for the wide deployment and acceptance of 3D-TV.

We have shown that visual attention is an important aspect to consider when addressing 3D human factors and that visual attention can be exploited to improve the quality of experience for 3D-TV program content. In particular, an understanding of 3D visual attention is important for the adaptation of depth for 3D content repurposing and for creating 3D content that can be viewed with greater visual comfort.

Recent results reported in the literature, including those by the authors [33], have indicated that locations of areas of interest may be different between the viewing of 2D and 3D stereoscopic content. In most cases, these differences are highly content-dependent. These results indicate that the introduction of disparity information may change the deployment of visual attention and that depth perception from 3D content plays an important role on our attentive behavior. Consequently, observations made from the presentation of 2D stimuli cannot be automatically generalized to 3D, and it is unlikely that models of 2D visual attention can be simply extended to 3D.

A simple extension of 2D visual attention models to two stereoscopic views is not really biologically plausible due to masking effects between views, occlusion or the effect of large disparities. More generally, it is argued that appropriate model architectures of 3D visual attention should consider the fusion of information from both eyes onto a single computational unit while keeping the correspondence of information from both eyes. Further extensive research is necessary concerning human behavioral responses in the visual exploration of 3D video content. More particularly, the role of each eye in the visual scanning and perception of 3D video content will require further investigation. Content is a strong influential factor in visual perception and further extensive studies are needed to fully understand the relationship between visual attention and features such as color, motion and depth.

Although research on 3D visual attention is still in its infancy both from the point of view of the recording 3D gaze in subjective experiments and from the point of view of the modeling, we have shown that considering 3D visual attention in the different processing steps of the 3D-TV delivery is important for creating an enjoyable 3D-TV viewing experience.

REFERENCES

[1] 3D@Home Consortium, “3D theater releases,” List available at http://www.3dathome.org/experience-theater.aspx, 2011.
[2] S. Jolly, M. Armstrong, and R. Salmon, “Three-dimensional television a broadcaster’s perspective,” in *Proc. SPIE Conf. Stereoscopic Displays and Applications XX*, vol. 7237, San Jose, January 2009.
[3] ISO/IEC JTC1/SC29/WG11 (MPEG) N10357 Video and Requirements Group, “Vision on 3D video,” February 2009.
[4] ITU-R Study Group 6, “Digital three-dimensional (3D) TV broadcasting,” Question ITU-R 128/6, 2008.

[5] ITU-T Study Group 9, "Objective and subjective methods for evaluating perceptual audiovisual quality in multimedia services within the terms of Study Group 9," Question 12/9, 2009.

[6] J. E. Cutting and P. M. Vishton, "Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth," in *Handbook of perception and cognition, Vol 5; Perception of space and motion*, W. Epstein and S. Rogers, Eds. San Diego: Academic Press, 1995, pp. 69–117.

[7] K. Hamburger, T. Hansen, and K. R. Gegenfurtner, "Geometric-optical illusions at isoluminance," *Vision Research*, vol. 47, no. 26, pp. 3276–3285, December 2007.

[8] E. Svarverud, S. J. Gilson, and A. Glennerster, "Cue combination for 3D location judgements," *Journal of Vision*, vol. 10, no. 1, pp. 1–13, January 2010.

[9] G. P. Bingham, A. Bradley, M. Bailey, and R. Vinner, "Accommodation, occlusion, and disparity matching are used to guide reaching: A comparison of actual versus virtual environments," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 27, no. 6, pp. 1314–1334, December 2001.

[10] Y. Okada, K. Ukai, J. S. Wolffsohn, B. Gilmartin, A. Iijima, and T. Bando, "Target spatial frequency determines the response to conflicting defocus-and convergence-driven accommodative stimuli," *Vision Research*, vol. 46, no. 4, pp. 475–484, February 2006.

[11] D. M. Hoffman, A. R. Girshick, K. Akeley, and M. S. Banks, "Vergence accommodation conflicts hinder visual performance and cause visual fatigue," *Journal of Vision*, vol. 8, no. 3, pp. 1–30, 2008.

[12] W. Chen, J. Fournier, M. Barkowsky, and P. Le Callet, "New requirements of subjective video quality assessment methodologies for 3D-TV," in *Proc. Fifth Int. Workshop on Video Processing and Quality Metrics (VPQM)*, Scottsdale, January 2010.

[13] D. A. Palmer, "Measurement of the horizontal extent of Panum's area by a method of constant stimuli," *Optica Acta: International Journal of Optics*, vol. 8, no. 2, pp. 151–159, 1961.

[14] D. E. Mitchell, "Retinal disparity and diplopia," *Vision Research*, vol. 6, no. 7-8, pp. 441–451, August 1966.

[15] I. P. Howard and B. J. Rogers, "Binocular fusion and rivalry," in *Seeing in depth, Vol. 1, Basic mechanisms*. I. Porteous, 2002, pp. 271–315.

[16] L. M. Wilcox, J. M. Harris, and S. P. McKee, "The role of binocular stereopsis in monoptic depth perception," *Vision Research*, vol. 47, no. 18, pp. 2367–2377, August 2007.

[17] A. L. Yarbus, *Eye movements and vision*. New-York: Plenum Press, 1967.

[18] M. I. Posner, "Orienting of attention," *Quaterly Journal of Experimental Psychology*, vol. 32, no. 1, pp. 3–25, 1980.

[19] A. M. Treisman and G. Gelade, "Feature integration theory of attention," *Cognitive Psychology*, vol. 12, pp. 97–136, January 1980.

[20] C. Koch and S. Ullman, "Shifts in selection in visual attention: Towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, no. 4, pp. 219–227, 1985.

[21] O. Le Meur and P. Le Callet, "What we see is most likely to be what matters: visual attention and applications," in *Proc. IEEE Int. Conf. Image Processing*, Cairo, November 2009, pp. 3085–3088.

[22] L. Itti, C. Koch, and E. Niebur, "Model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, November 1998.

[23] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 802–819, May 2006.

[24] N. D. B. Bruce and J. K. Tsotsos, "Saliency based on information maximization," *Advances in Neural Information Processing Systems*, vol. 18, pp. 155–162, June 2006.

[25] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vision Research*, vol. 47, 2007.

[26] X. Hou and L. Zhang, "Saliency detection: a spectral residual approach," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 17, no. 22, Minneapolis, June 2007, pp. 1–8.

[27] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A Bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, pp. 1–20, 2008.

[28] L. Itti and P. F. Baldi, "Bayesian surprise attracts human attention," *Vision Research*, vol. 49, no. 10, pp. 1295–1306, May 2009.

[29] L. Jansen, S. Onat, and P. König, "Influence of disparity on fixation and saccades in free viewing of natural scenes," *Journal of Vision*, vol. 9, no. 1, pp. 1–19, January 2009.

[30] D. A. Wismeijer, C. J. Erkelens, R. van Ee, and M. Wexler, "Depth cue combination in spontaneous eye movements," *Journal of Vision*, vol. 10, no. 6, pp. 1–15, June 2010.

[31] J. Häkkinen, T. Kawai, J. Takatalo, R. Mitsuya, and G. Nyman, "What do people look at when they watch stereoscopic movies?" in *Proc. SPIE Conf. Stereoscopic Displays and Applications XXI*, vol. 7524, San Jose, January 2010.

[32] C. Ramasamy, D. House, A. Duchowski, and B. Daugherty, "Using eye tracking to analyze stereoscopic filmmaking," in *Proc. SIGGRAPH 2009:Posters*, 2010.

[33] Q. Huynh-Thu and L. Schiatti, "Examination of 3D visual attention in stereoscopic video content," in *Proc. SPIE Conference on Human Vision and Electronic Imaging XVI*, San Francisco, January 2010.

[34] A. Maki, J. O. Eklundh, and P. Nordlund, "A computational model of depth-based attention," in *Proc. Int. Conf. Pattern Recognition*, vol. 4, Vienna, August 1996.

[35] A. Maki, P. Nordlund, and J. O. Eklundh, "Attentional scene segmentation: Integrating depth and motion from phase," *Computer Vision and Image Understanding*, vol. 78, pp. 351–373, 2000.

[36] N. Ouerhani and H. Hügli, "Computing visual attention from scene depth," in *Proc. Int. Conf. Pattern Recognition*, Barcelona, 2000, pp. 375–378.

[37] N. D. B. Bruce and J. K. Tsotsos, "An attentional framework for stereo vision," in *Proc. 2nd Canadian Conf. Computer and Robot Vision*, Victoria, May 2005, pp. 88–95.

[38] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artificial Intelligence*, vol. 78, no. 1-2, pp. 507–545, 1995.

[39] Y. Zhang, G. Jiang, M. Yu, and K. Chen, "Stereoscopic visual attention model for 3D video," in *Advances in Multimedia Modeling*, ser. Lecture Notes in Computer Science. Berlin: Springer, 2010, vol. 5916, pp. 314–324.

[40] K. Essig, M. Pomplun, and H. Ritter, "A neural network for 3D gaze recording with binocular eye trackers," *Int. Journal of Parallel, Emergent and Distributed Systems*, vol. 21, no. 2, pp. 79–95, April 2006.

[41] Y.-M. Kwon, K.-W. Jeon, J. Ki, Q. M. Shahab, S. Jo, and S.-K. Kim, "3D gaze estimation and interaction to stereo display," *Int. Journal of Virtual Reality*, vol. 5, no. 3, pp. 41–45, September 2006.

[42] J. Chen, Y. Tong, W. Gray, and Q. Ji, "A robust 3D eye gaze tracking system using noise reduction," in *Proc. 2008 symposium on eye tracking research applications (ETRA)*, vol. 1, March 2008, pp. 189–196.

[43] J. Kim, A. Baik, Y. J. Jung, and D. Park, "2D-to-3D conversion by using visual attention analysis," in *Proc. SPIE Conf. Stereoscopic Displays and Applications XXI*, vol. 7524, San Jose, January 2010.

[44] C. Chamaret and O. Le Meur, "Attention-based video reframing: validation using eye-tracking," in *Proc. Int. Conf. Pattern Recognition*, Tampa, December 2008, pp. 1–4.

[45] X. Fan, X. Xie, W.-Y. Ma, H.-J. Zhang, and H.-Q. Zhou, "Visual attention based image browsing on mobile devices," in *Proc. Int. Conf. on Multimedia and Expo*, vol. 2, 2003, pp. 53–56.

[46] O. Le Meur, S. Cloarec, and P. Guillotel, "Automatic content repurposing for mobile applications," *SMPTE Motion Imaging Journal*, vol. 119, no. 1, pp. 30–34, February 2010.

[47] C. Chamaret, S. Godeffroy, P. Lopez, and O. Le Meur, "Adaptive 3D rendering based on region-of-interest," in *Proc. SPIE Conf. Stereoscopic Displays and Applications XXI*, vol. 7524, San Jose, January 2010.

[48] ITU-T Study Group 16, "Advanced video coding for generic audiovisual services," ITU-T Recommendation H.264, March 2005.

[49] Z. Li and L. Itti, "Visual attention guided video compression," in *Proc. Vision Science Society Annual Meeting*, Naples, May 2008.

[50] M. Barkowsky, K. Wang, R. Cousseau, K. Brunnström, R. Olsson, and P. Le Callet, "Subjective quality assessment of error concealment strategies for 3D-TV in the presence of asymmetric transmission errors," in *Proc. Packet Video Workshop*, Hong Kong, December 2010.

[51] J. Carreira, L. Pinto, N. Rodrigues, S. Faria, and P. Assuncao, "Subjective assessment of frame loss concealment methods in 3D video," in *Proc. Picture Coding Symposium*, Nagoya, December 2010, pp. 182– 185.
[52] S. Pastoor, "Human factors of 3D displays in advanced image communications," *Displays*, vol. 14, no. 3, pp. 150–157, July 1993.
[53] S. Nagata, "The binocular fusion of human vision on stereoscopic displays-field of view and environment effects," *Ergonomics*, vol. 39, no. 11, pp. 1273–1284, November 1996.
[54] M. Wöpking, "Viewing comfort with stereoscopic pictures: An experiment study on subjective effects of disparity magnitude and depth of focus," *Journal of the Society for Information Display*, vol. 3, no. 3, pp. 101–103, 1995.
[55] J. J. Semmlow and D. Heerema, "The role of accommodative convergence at the limit of fusional vergence," *Invest. Ophthalmology and Visual Science*, vol. 18, pp. 970–976, 1979.
[56] K. Talmi and J. Liu, "Eye and gaze tracking for visually controlled interactive stereoscopic displays," *Signal Processing Image Communication.*, vol. 14, no. 10, pp. 799–810, August 1999.
[57] M. Lambooij, W. IJsselsteijn, M. Fortuin, and I. Heynderickx, "Visual discomfort and visual fatigue of stereoscopic displays: A review," *Journal of Imaging Science and Technology*, vol. 53, no. 3, pp. 1–14, May 2009.
[58] G. Mather and D. R. R. Smith, "Depth cue integration: stereopsis and image blur," *Vision Research*, vol. 40, no. 25, pp. 3501–3506, January 2000.
[59] S. Hillaire, A. Lécuyer, R. Cozot, and G. Casiez, "Depth-of-field blur effects for first-person navigation in virtual environments," *IEEE Computer Graphics and Application*, vol. 28, no. 6, pp. 47–55, November 2008.
[60] ——, "Using an eye-tracking system to improve depth-of-field blur effects and camera motions in virtual environments," in *Proc. IEEE Virtual Reality*, Reno, March 2008, pp. 47–51.
[61] S. Hillaire, A. Lécuyer, G. Breton, and T. R. Corte, "Gaze behavior and visual attention model when turning in virtual environments," in *Proc. ACM Symposium on Virtual Reality Software and Technology*, Kyoto, November 2009, pp. 43–50.

Quan Huynh-Thu holds the Dipl.-Ing. degree in electrical engineering from the University of Liège (Belgium), the M.Eng. degree in electronics engineering from the University of Electro-Communications (Japan) and the Ph.D. degree in electronic systems engineering from the University of Essex (UK).

Dr. Huynh-Thu is currently Senior Scientist at Technicolor Research & Innovation, France. Prior to that, he was Research Scientist in the Image and Signal Processing Lab at the Belgian Forensic Institute from 1997 to 2000. He was awarded a postgraduate fellowship from the Japanese Ministry of Education and was Researcher at the University of Electro-Communications, Tokyo, from 2000 to 2003. He was then Senior Research Engineer at Psytechnics Ltd (UK) from 2003 to 2010, where he conducted research on perceptual video quality and co-developed the ITU-T Recommendation J.247 for the objective measurement of perceptual video quality using a full-reference model.

His current research interests include 3D human factors, visual perception, visual attention, and video quality assessment. He has been actively contributing to the work of the International Telecommunication Union (ITU) and the Video Quality Experts Group (VQEG) since 2004. He is currently co-chair of the VQEG 3D-TV and Multimedia groups, and Rapporteur for Qs2 in ITU-T Study Group 9.

Marcus Barkowsky received his Dipl.-Ing. degree in Electrical Engineering from the University of Erlangen-Nuremberg, Germany, in 1999.

Starting from a deep knowledge of video coding algorithms his Ph.D. thesis focused on a reliable video quality measure for low bitrate scenarios. Special emphasis on mobile transmission led to the introduction of a visual quality measurement framework for combined spatio-temporal processing with special emphasis on the influence of transmission errors.

He received the Dr.-Ing. degree from the University of Erlangen-Nuremberg in 2009. Since November 2008, he is researching the relationship between the human visual system and the technological issues of 3D television at the University of Nantes, France. His current activities range from modeling the influence of coding, transmission, and display artifacts in 2D and 3D to measuring and quantifying visual discomfort and visual fatigue on 3D displays.

Patrick Le Callet is currently a Full Professor at Ecole Polytechnique de l'Université de Nantes. He has been teaching as an Assistant Professor from 1997 to 1999 and as a full-time Lecturer from 1999 to 2003 in the Department of Electrical Engineering, Technical Institute of University of Nantes (IUT). Since 2003, he has been teaching at Ecole Polytechnique de l'Université de Nantes (Engineering School) in the Electrical Engineering and the Computer Science Departments. In 1997, he joined the Image and Video Communication group at CNRS IRCCyN. Since 2006, he has been the head of this group that includes ten permanent professors, two assistant professors, 18 postdoctoral and Ph.D. students, and five research engineers. His research focuses on better understanding of the human visual system and designing and applying HVS models in image and video processing. Current topics of interest are DTV, image, and video quality assessment, watermarking techniques, and visual attention modeling and applications. He is co-author of more than 70 publications/communications and co-recipient of six international patents.